

REPORT DOCUMENTATION PAGE			Form Approved OMB NO. 0704-0188		
<p>The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA, 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.</p> <p>PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.</p>					
1. REPORT DATE (DD-MM-YYYY) 01-03-2016		2. REPORT TYPE Final Report		3. DATES COVERED (From - To) 24-Aug-2011 - 23-Aug-2012	
4. TITLE AND SUBTITLE Final Report on "Survey of Quantification and Distance Functions Used for Internet-based Weak-link Sociological Phenomena"			5a. CONTRACT NUMBER W911NF-11-1-0322		
			5b. GRANT NUMBER		
			5c. PROGRAM ELEMENT NUMBER 611102		
6. AUTHORS Ming-Jun Lai, Dawn Robinson			5d. PROJECT NUMBER		
			5e. TASK NUMBER		
			5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAMES AND ADDRESSES University of Georgia Research Foundation, 200 D. W. Brooks Drive Athens, GA 30602 -5016			8. PERFORMING ORGANIZATION REPORT NUMBER		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS (ES) U.S. Army Research Office P.O. Box 12211 Research Triangle Park, NC 27709-2211			10. SPONSOR/MONITOR'S ACRONYM(S) ARO		
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) 60673-MA.4		
12. DISTRIBUTION AVAILABILITY STATEMENT Approved for Public Release; Distribution Unlimited					
13. SUPPLEMENTARY NOTES The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.					
14. ABSTRACT The PI studied all mathematical literature he can find related to the Google search engine, Google matrix, PageRank as well as the Yahoo search engine and a classic SearchKing HIST algorithm. The co-PI immersed herself in the sociology literature for the relevant studies on social network, strong and weak social ties and measures of centrality, etc.. which will be useful for the study of search algorithms. Our findings are summarized in three separate manuscripts attached.					
15. SUBJECT TERMS Google Matrix, PageRank, Bonacich Centrality Measure, In- and Out-Linkages, Damping Factor, Secondary Linkage					
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT UU	15. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON Ming-Jun Lai
a. REPORT UU	b. ABSTRACT UU	c. THIS PAGE UU			19b. TELEPHONE NUMBER 706-542-2065

Report Title

Final Report on "Survey of Quantification and Distance Functions
Used for Internet-based Weak-link Sociological Phenomena"

ABSTRACT

The PI studied all mathematical literature he can find related to the Google search engine, Google matrix, PageRank as well as the Yahoo search engine and a classic SearchKing HIST algorithm. The co-PI immersed herself in the sociology literature for the relevant studies on social network, strong and weak social ties and measures of centrality, etc.. which will be useful for the study of search algorithms. Our findings are summarized in three separate manuscripts attached.

Enter List of papers submitted or published that acknowledge ARO support from the start of the project to the date of this printing. List the papers, including journal references, in the following categories:

(a) Papers published in peer-reviewed journals (N/A for none)

Received

Paper

TOTAL:

Number of Papers published in peer-reviewed journals:

(b) Papers published in non-peer-reviewed journals (N/A for none)

Received

Paper

TOTAL:

Number of Papers published in non peer-reviewed journals:

(c) Presentations

Number of Presentations: 0.00

Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Non Peer-Reviewed Conference Proceeding publications (other than abstracts):

Peer-Reviewed Conference Proceeding publications (other than abstracts):

Received Paper

TOTAL:

Number of Peer-Reviewed Conference Proceeding publications (other than abstracts):

(d) Manuscripts

Received Paper

TOTAL:

Number of Manuscripts:

Books

Received Book

TOTAL:

Received Book Chapter

TOTAL:

Patents Submitted

Patents Awarded

Awards

Graduate Students

<u>NAME</u>	<u>PERCENT_SUPPORTED</u>
FTE Equivalent:	
Total Number:	

Names of Post Doctorates

<u>NAME</u>	<u>PERCENT_SUPPORTED</u>
FTE Equivalent:	
Total Number:	

Names of Faculty Supported

NAME

PERCENT SUPPORTED

FTE Equivalent:

Total Number:

Names of Under Graduate students supported

NAME

PERCENT SUPPORTED

FTE Equivalent:

Total Number:

Student Metrics

This section only applies to graduating undergraduates supported by this agreement in this reporting period

The number of undergraduates funded by this agreement who graduated during this period: 0.00

The number of undergraduates funded by this agreement who graduated during this period with a degree in science, mathematics, engineering, or technology fields:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and will continue to pursue a graduate or Ph.D. degree in science, mathematics, engineering, or technology fields:..... 0.00

Number of graduating undergraduates who achieved a 3.5 GPA to 4.0 (4.0 max scale):..... 0.00

Number of graduating undergraduates funded by a DoD funded Center of Excellence grant for Education, Research and Engineering:..... 0.00

The number of undergraduates funded by your agreement who graduated during this period and intend to work for the Department of Defense 0.00

The number of undergraduates funded by your agreement who graduated during this period and will receive scholarships or fellowships for further studies in science, mathematics, engineering or technology fields: 0.00

Names of Personnel receiving masters degrees

NAME

Total Number:

Names of personnel receiving PHDs

NAME

Total Number:

Names of other research staff

NAME

PERCENT SUPPORTED

FTE Equivalent:

Total Number:

Sub Contractors (DD882)

Inventions (DD882)

Scientific Progress

The PI studied all mathematical literature he can find related to the Google search engine, Google matrix, PageRank as well as the Yahoo search engine and a classic SearchKing HIST algorithm. The co-PI immersed herself in the sociology literature for the relevant studies on social network, strong and weak social ties and etc.. which may be useful for the study of search algorithms. Our findings can be summarized as follows.

{1}} A review of the related mathematical literature is given in section 2.2 of Appendix I.

{2}} In addition, the PI presented a new proof of the convergence of Google searching algorithm although there are several proofs available already. See Section 2.3 of Appendix I. This analysis is used for new algorithms based on sociological analysis discussed in 6) below.

{3}} The PI proposed a local update algorithm which is new. However, its computational efficiency is not compared with other updated algorithms yet. See Section 2.4 of Appendix I.

{4}} Certainly, the PI studied how to improve the efficiency of the computation in hoping to speed up the search. Due to the explosive increase of the webpages and the Internet surfers, such a study is definitely necessary. However, the study does not lead to any meaningful results so far.

{5}} The PI discussed with the co-PI about the web search algorithms in sociological sense which is one of the key points of this project. The PI found that the Bonacich centrality measure is the same as the one used in Google search algorithm if the relation matrix is the adjacency matrix with an appropriate normalization constant. See section 3 of Appendix I.

{6}} The co-PI suggested to use not only the out-linkages, but also the in-linkages in the search algorithm. In addition, the co-PI suggested to use the secondary linkages to further help determine the relevance in PageRank. Based on these ideas, the PI has proposed a new version of searching algorithm which uses both in- and out-linkages, and proved that the modified algorithm is still convergent and the convergence rate is determined by the damping factor. See section 4 of Appendix I.

{7}} The co-PI reviewed a lot of literature in sociology related to the PageRank, the relevant scores for linkages, in particular, for social linkages, as well as the social strong and weak ties, social cohesion, position and distance. See Appendix II.

{8}} The co-PI was willing to write it out a section to explain why one needs to use both in- and out-linkages in sociological terminology. However, the PI has not received such a write-up as Jan. 15, 2013.

{9}} Finally, the PI studied the Yahoo search algorithm based on its patent although the current Yahoo updates its search algorithm and the PI did not find enough time to study it. See Section 5 of Appendix I.

Appendix I:

Ming-Jun Lai and Dawn Robinson, A Mathematical and Sociological Analysis of Google Search Algorithm, a technical report, 2013.

Appendix II:

Dawn Robinson, A Summary of Sociological Concepts Related to Social Network and Its Techniques for Quantifying Social Cohesion, Social Position, Social Distance, a technical report, 2013.

Technology Transfer

Final Report on "Survey of Quantification and Distance Functions Used for Internet-based Weak-link Sociological Phenomena" *

Ming-Jun Lai [†]

Dawn Robinson[‡]

January 16, 2013

1 Research Activities:

Dr. Ming-Jun Lai, the PI and Dr. Dawn Robinson, the co-PI had a few meetings in the Lai's office and one meeting at the Robinson's office and discussed how to proceed the research during the Aug. and Sept., 2011. They had several emails communications before and after these meeting. Also, later in March 2012, they had another meeting on the distribution of the grant money, one summer month salary for the PI and one summer month salary for the co-PI. The PI used all the fund allocated for travel, a part of the travel fund was used for Mr. George Slavov, one of the PI's graduate students for his summer salary and a part of the travel fund was used for Mr. Leopold Matamba-Messi's travel to UCLA for a conference. Mr. Leopold Matamba Messi is another graduate students of the PI. Mr. Matamba-Messi graduated in the Aug. 2012 and found a post-doc position at the Mathematical Biology Institute at Ohio State University. Since Aug., 2012, the PI has emailed to the co-PI about the final report several times. The PI and co-PI met again on Dec. 5, 2012 during a workshop and discussed how to write this report. They have met again in the Lai's office on Jan. 16, 2013 for submission of the final report.

2 Research Results:

The PI studied all mathematical literature he can find related to the Google search engine, Google matrix, PageRank as well as the Yahoo search engine and a classic SearchKing HIST algorithm. The co-PI immersed herself in the sociology literature for the relevant studies on social network, strong and weak social ties and etc.. which may be useful for the study of search algorithms. Our findings can be summarized as follows.

- 1) A review of the related mathematical literature is given in section 2.2 of Appendix I.
- 2) In addition, the PI presented a new proof of the convergence of Google searching algorithm although there are several proofs available already. See Section 2.3 of Appendix I. This analysis is used for new algorithms based on sociological analysis discussed in 6) below.

*This project is supported by Army Research Office Grant # W911NF-11-1-0322, August 23, 2011–Aug. 23, 2012

[†]Dept. of Mathematics, University of Georgia, Athens, GA 30602. Email Address: mjlai@math.uga.edu

[‡]Dept. of Sociology, University of Georgia, Athens, GA 30602. Email Address: mjlai@math.uga.edu. Email address: sodawn@uga.edu

- 3) The PI proposed a local update algorithm which is new. However, its computational efficiency is not compared with other updated algorithms yet. See Section 2.4 of Appendix I.
- 4) Certainly, the PI studied how to improve the efficiency of the computation in hoping to speed up the search. Due to the explosive increase of the webpages and the Internet surfers, such a study is definitely necessary. However, the study does not lead to any meaningful results so far.
- 5) The PI discussed with the co-PI about the web search algorithms in sociological sense which is one of the key points of this project. The PI found that the Bonacich centrality measure is the same as the one used in Google search algorithm if the relation matrix is the adjacency matrix with an appropriate normalization constant. See section 3 of Appendix I.
- 6) The co-PI suggested to use not only the out-linkages, but also the in-linkages in the search algorithm. In addition, the co-PI suggested to use the secondary linkages to further help determine the relevance in PageRank. Based on these ideas, the PI has proposed a new version of searching algorithm which uses both in- and out-linkages, and proved that the modified algorithm is still convergent and the convergence rate is determined by the damping factor. See section 4 of Appendix I.
- 7) The co-PI reviewed a lot of literature in sociology related to the PageRank, the relevant scores for linkages, in particular, for social linkages, as well as the social strong and weak ties, social cohesion, position and distance. See Appendix II.
- 8) The co-PI was willing to write it out a section to explain why one needs to use both in- and out-linkages in sociological terminology. However, the PI has not received such a write-up as Jan. 15, 2013.
- 9) Finally, the PI studied the Yahoo search algorithm based on its patent although the current Yahoo updates its search algorithm and the PI did not find enough time to study it. See Section 5 of Appendix I.

3 Conclusion:

The PI has surveyed all the mathematical literature related to the search engines up to today and the co-PI has summarized all the sociological concepts related to the internet search, not only the relevance among webpages, but also the closeness, betweenness among people, groups, etc.. The goal of this project to understand the PageRank sociologically and how it could be adjusted to be more effective sociologically as well as how to make the algorithms more efficient. This is still a central issue. The PI is not able to complete this goal of the project as the PI and co-PI have managed their time to only do the surveys of the current literatures in mathematical and sociological senses and have not found enough time to discuss their connections. They made a small progress towards the goal as mentioned above. The PI would like to continue this project if the ARO is willing to fund it again. Due to this project, the PI gains an excellent knowledge on the mathematical study of the Google searching algorithm and other search algorithms. He is ready to attack some research problems related to the algorithms such as how to speed up the computation and how to update in parallel the Google matrix, its PageRank vector and etc.. Also, he is now able to work with a sociologist better as he knows many sociological definitions and notations. See Items 5) and 6) above as an evidence.

With more time, this research team could more deeply digest the existing information, and propose some new algorithms, making use of sociological insights to improve mathematically the characterization of relations in large, complex systems. The PI and co-PI plan to continue their joint work toward a better understanding of the searching algorithms and discovering how to better situate them in the larger contexts of existing mathematical and sociological knowledge.

Appendix I

Ming-Jun Lai and Dawn Robinson, A Mathematical and Sociological Analysis of Google Search Algorithm, a manuscript, 2013.

Appendix II

Dawn Robinson, A Summary of Sociological Concepts Related to Social Network and Its Techniques for Quantifying Social Cohesion, Social Position, Social Distance, a manuscript, 2013.

A Mathematical and Sociological Analysis of Google Search Algorithm *

Ming-Jun Lai[†]

Dawn Robinson[‡]

January 16, 2013

Abstract

Google search algorithm for finding relevant information on-line based on keywords, phrases, links, and webpages is analyzed in the mathematical and sociological settings in this article. We shall first survey mathematical study related to the Google search engine and then present a new analysis for the convergence of the search algorithm and a new update scheme. Next based on sociological knowledge, we propose to use in- and out- linkages as well as use the second order linkages to refine and improve the search algorithm. We use the sociology to justify our proposed improvements and mathematically prove the convergence of these two new search algorithms.

1 Introduction

Due the huge volume of documents available through the world wide web, a search engine or similar service is now absolutely necessary in order to find relevant information on-line. About 12 years ago, Google, Yahoo and other companies started providing such a service. It is interesting to know the mathematics of computational algorithms behind these search engines, in particular, the Google search engine which is so successful nowadays. Many mathematicians and computer scientists, G. Golub and his collaborators, Brezinski and his collaborators, Langville and Meyer, among other pioneers worked on the mathematics of the Google matrix and its computation. See [25], [16], [9], [8], [31], [32] and a recent survey in [2]. The search algorithm also becomes a very useful tool in many Web search technologies such as spam detection, crawler configuration, trust networks, and etc.. and find many applications in [12], [13] and etc.. Today, it is the most used computational algorithm in the world. More and more websites such as Facebook, Amazon, Netflix, and etc.. use similar search algorithms for various purposes.

As almost all people on the earth use the Google search once or several times daily, it is interesting to see if the search results make any sense in sociology. How could it be adjusted to be more effective sociologically. We divide this article into two parts. We shall first present a mathematical description of the computational algorithm behind the Google search engine and summarize the recent studies related to the algorithm. Then we present a sociological analysis. Based on the sociology, we propose two modifications in order to be more effective.

*This research is supported by Army Research Office Grant # W911NF-11-1-0322.

[†]mjlai@math.uga.edu. Department of Mathematics, The University of Georgia, Athens, GA 30602.

[‡]sodawn@uga.edu. Department of Sociology, The University of Georgia, Athens, GA 30602.

2 Mathematical Analysis

In this section we first survey the mathematical studies related to the Google search algorithm. Then we present a mathematical analysis which is new to the best of our knowledge. Mainly we study its convergence as it is an iterative method since the Google matrix is of huge size so that any direct method is impossible. Then we shall establish the convergence, present the rate of convergence and a local updating scheme.

2.1 Google Search Engine

During each search, after entering a key word or key words, the Google search engine will produce a ranking vector called PageRank. It is a vector of huge size (several billion entries) although almost all entries are zero in general. Each entry of the vector presents a webpage. Nonzero value in an entry is a weighting factor deciding how important the page related to the key word(s) just entered. The importance is determined by popularity. The ideas for this computational algorithm have been reported as early as 1997–1998 by Page in [37] and Brin and Page in [10]. See also [33] for explanation of the ideas. The algorithm mainly computes a numerical weighting (called pagerank) to each webpage or each element in a hyperlinked set of documents. Pageranks reflect the Google's view of the importance of webpages among more than several billion webpages. It is a pragmatic approach to improve search quality by going through the collective intelligence of the web to determine a page's importance.

Let \mathbf{v} be a vector of \mathbb{R}^N with $N \geq 8 \text{ billion}$. Any unit vector in \mathbb{R}^N is associated with a webpage. Mathematically, Pagerank is a probability function of vectors. For a page \mathbf{v} , let $P(\mathbf{v})$ be the PageRank of \mathbf{v} . By a stroke of the ingenious, Brin and Page thought that $P(\mathbf{v})$ is the weighted sum of the PageRanks of all pages pointing to \mathbf{v} . That is,

$$P(\mathbf{v}) = \sum_{\mathbf{u} \in A_{\mathbf{v}}} P(\mathbf{u}) \ell(\mathbf{u}, \mathbf{v}), \quad (1)$$

where $A_{\mathbf{v}}$ is the collection of all pages connected to \mathbf{v} and $\ell(\mathbf{u}, \mathbf{v}) \in [0, 1]$ be a weight dependent on the number of links from page \mathbf{u} to the other pages. More precisely, let $L(\mathbf{u})$ be the number of linkages from \mathbf{u} and then define $\ell(\mathbf{u}, \mathbf{v}) = 1/L(\mathbf{u})$ if $\mathbf{u} \in A_{\mathbf{v}}$ and 0 otherwise for any other page \mathbf{u} . Such a definition is cyclic. Also, a web surfer may get bored and abandons the current search by teleporting to a new page. By a second stroke of ingenious, Brin and Page introduced a damping factor and modified the Pagerank function $P(\mathbf{v})$ to be

$$P(\mathbf{v}) = d \sum_{\mathbf{u} \in A_{\mathbf{v}}} P(\mathbf{u}) \ell(\mathbf{u}, \mathbf{v}) + \frac{1-d}{N}, \quad \forall \text{unit vector } \mathbf{v} \in \mathbb{R}^N, \quad (2)$$

where $d \in (0, 1)$ is a damping factor. It indeed leads to a convergent algorithm and the PageRank P will be unique defined. In summary, we have

- PageRank

$$\mathbf{P} = [P(\mathbf{v}_1), \dots, P(\mathbf{v}_N)]^T. \quad (3)$$

- $M = [\ell(\mathbf{v}_i, \mathbf{v}_j)]_{1 \leq i, j \leq N}$ be the adjacency matrix of size $N \times N$ with convention $\ell(\mathbf{v}_i, \mathbf{v}_i) = 0$.
- $\mathbf{1} = [1, 1, \dots, 1]^T \in \mathbb{R}^N$;

- The Google Matrix

$$\mathcal{G} = dM + \frac{1-d}{N}\mathbf{1} \cdot \mathbf{1}^\top. \quad (4)$$

which is a stochastic matrix, where $d > 0$ is a damping factor. In Google search engine, $d = 0.85$.

To compute the PageRank \mathbf{P} , the Google search engine computes: starting with an initial vector \mathbf{P}_1 , e.g., $\mathbf{P}_1 = \mathbf{1}/N$,

$$\mathbf{P}_{k+1} = \mathcal{G}\mathbf{P}_k, \forall k = 1, 2, \dots \quad (5)$$

until the difference of \mathbf{P}_{k+1} and \mathbf{P}_k within a given tolerance.

For an elementary explanation of Google computational algorithm, see a monograph [33]. Many mathematical problems arise from the algorithm. For example, does the computation (5) converge and what is the convergence rate? Is the computation stable? How to speed up the iterations in (5)? How to do the computation in parallel? Can we have $d = 1$? How can we add or delete a page web from the PageRank function?

2.2 Summary of Mathematical Results

The major question for the computational algorithm (5) in the previous subsection is the convergence rate of the iterations in (5). The iteration is a power method. The convergence is dependent on the second largest eigenvalue as the first largest eigenvalue is 1 (cf. [27]). The value of the second largest eigenvalue or eigenvalues is the damping factor as shown first in [24]. This fact is later established in [39] by using Jordan canonical form of the Google matrix. It was also proved in [14] where the researcher showed that the second largest eigenvalue is the damping factor even the Google matrix has more than 1 largest eigenvalues. It is also proved in [30]. In [11], the searcher interpreted the PageRank algorithm as a power series which leads to a new proof of the several known results about the PageRank, e.g., the convergence rate is the exactly the damping factor by [24] and presented a new inside about PageRank which is the probability that a specific type of random surfer ends his walk at the page. Usually, the random surfer is assumed to be uniform distribution suggested by Page and Brin. In fact, all the properties of the PageRank hold for arbitrary probability setting. We shall present another proof in the next subsection

A second question is if the computation is stable or what the condition number of the Google matrix is. In [18], the researchers viewed the PageRank vector to the stationary distribution of a stochastic matrix, the Google matrix. They analyzed the sensitivity of PageRank to changes in the Google matrix, including addition and deletion of links in the web graph and presented error bounds for the iterates of the power method and for their residuals. See also various properties in [9] including the condition number of the Google matrix.

A central question is how to accelerate the Google computational algorithm. There are several approaches.

- *Extrapolation* In [8], the researchers studied the computation of the nonnegative left eigenvector associated with the dominant eigenvalue of the PageRank matrix and proposed to use extrapolation methods to approximate the eigenvector. Furthermore, in [9], the researchers proposed some acceleration methods due to the slow convergence of the Power method for the PageRank vector. The acceleration method can be explained as the method of Vorobyev moments and Krylov subspace method. In addition, many properties about the Google matrix and PageRank vectors were presented.

- *Sparsity Split* In [40], these Chinese researchers use the sparsity of the hyperlink matrix (one of the two matrices in Google Matrix) and proposed a linear extrapolation to optimize each iteration in the Power method for the computation of PageRank and reduction of the storage space. The idea is simple and the engineers in Google may have already implemented.
- *Iterative Solutions* The iteration in (5) can be reinterpreted as a linear system to be given in (8). As explained in [15], the Google matrix is very large, sparse and non-symmetric. Solution of the linear system by a direct method is not feasible due to the matrix size. Many iterative solutions such as Gauss-Jacobi iterations, Generalize Minimum Residual (GMRES), Biconjugate Gradient (BiCG), Quasi-Minimal Residual (QMR), Conjugate Gradient Squared (CGS), Biconjugate Gradient Stabilized (BiCGSTAB), Chebyshev Iterations. The researchers proposed to use the parallel Block Jacobi iteration with adaptive Schwarz preconditioners.
- *Parallel Iteration* In [15], the researchers described a parallel implementation for PageRank computation and demonstrated that the parallel PageRank computing is faster and less sensitive to the changes in teleportation.
- *Lumping Dangling Nodes* In [19], the researchers lumped all of the dangling nodes together into a single node and the Google matrix can be reduced. They showed that the reduced stochastic matrix has the same nonzero eigenvalues as the full Google matrix and the convergence rate is the same when the Power method is applied. In [34], the researchers further reduced the Google matrix by lumping more nodes which are so-called weakly nondangling nodes together and the reduced matrix has the same nonzero eigenvalues as the Google matrix.
- *Distributed Randomized Algorithm* In a series of papers ([20], [21], [22], the researchers applied distributed randomized algorithm for PageRank. The idea is to let each page compute its own value by exchanging information with its linked pages. The communication among the pages is randomized so as to make it asynchronous. The communication among pages in the sense that data transmitted over the links is received correctly or could have some noises.

There are several other improvements. For example,

One is to update the PageRank. In [33], the researchers considered the PageRank vector to be an state of homogeneous irreducible Markov chain and the chain requires updating by altering some of its transition probabilities or by adding or deleting some states. They proposed a general purpose algorithm which simultaneously deals with both kinds of updating problems and proved its convergence. We shall present another approach for local updating to be discussed in the next subsection.

Another improvement is to distinct the webpage with equal rankings. In [35], the researchers proposed an idea to organize the search result produced by the PageRank where several pages have the same rank score so that the surfer can get more relevant and important results easily. Their idea is to add a weight to each page.

We finally remark that besides the PageRank algorithm, there are many other search algorithms available. For example, Kleinberg's HITS algorithm is a classic Hypertext Induced Topics Search (HITS) algorithm invented by Kleinberg (cf. [28]). Kleinberg viewed that most pages are a hub and authority, i.e., a general page not only points to other pages and also receives points from other pages. Again, the HITS algorithm uses the power method to the hub weight vector and authority weight vector. See [33] and

[6] for more explanation. For another example, Yahoo uses another type of algorithms. See last section of this paper.

In summary, the study of Google computational algorithm requires many mathematical tools such that graph theory (cf. [?]), random walks and Markov chains (c.g., [4], [33]), numerical analysis (cf. [16]), [15], [8]), stochastic matrices (cf. [18]), networks theory (cf. [6]), etc.. See [2] for a survey of recent techniques for on link-based ranking methods and the PageRank algorithm to find the most relevant documents from the WWW.

2.3 Our Convergence Analysis

We see that $P(\mathbf{v}) \geq 0$. We now claim that $P(\mathbf{v}) \leq 1$. In fact, we can prove the following

Theorem 2.1 *Let $\mathbf{v}_i, i = 1, \dots, N$ be the standard unit vectors in \mathbb{R}^N . Then the PageRanks $P(\mathbf{v}_i), i = 1, \dots, N$ satisfy*

$$\sum_{i=1}^N P(\mathbf{v}_i) = 1.$$

Proof. Indeed, as in the definition (2), $P(\mathbf{v}_i), i = 1, \dots, N$ are dependent on each other. We recall

$$\mathbf{P} = [P(\mathbf{v}_1), \dots, P(\mathbf{v}_N)]^\top$$

is the PageRank vector in \mathbb{R}^N . We can see that

$$\sum_{i=1}^N \ell(\mathbf{v}_j, \mathbf{v}_i) = 1 \quad (6)$$

for $j = 1, \dots, N$. We call $\ell(\mathbf{v}_i, \mathbf{v}_j)$ adjacency functions. It follows from (2), we can write

$$\mathbf{P} = \frac{1-d}{N} \mathbf{1} + dM\mathbf{P}, \quad (7)$$

where $M = [\ell(\mathbf{v}_i, \mathbf{v}_j)]_{1 \leq i, j \leq N}$ be the adjacency matrix of size $N \times N$ with convention $\ell(\mathbf{v}_i, \mathbf{v}_i) = 0$ and $\mathbf{1} = [1, 1, \dots, 1]^\top \in \mathbb{R}^N$. Hence, we have

$$\begin{aligned} \sum_{i=1}^N P(\mathbf{v}_i) &= \mathbf{1}^\top \mathbf{P} = N \frac{1-d}{N} + d \mathbf{1}^\top M \mathbf{P} \\ &= 1 - d + d \mathbf{1}^\top \mathbf{P} = 1 - d + d \sum_{i=1}^N P(\mathbf{v}_i), \end{aligned}$$

where we have used (6) to have $\mathbf{1}^\top M = \mathbf{1}^\top$. That is, we have $(1-d) \sum_{i=1}^N P(\mathbf{v}_i) = 1 - d$ and hence,

$\sum_{i=1}^N P(\mathbf{v}_i) = 1$ since $d \neq 1$. This completes the proof. ■

The PageRank algorithm offers two ways to calculate the PageRank vector \mathbf{P} : one is by algebraic method and the other by an iterative method. Indeed, by (7), we have

$$(I - dM)\mathbf{P} = \frac{1-d}{N}\mathbf{1} \quad \text{or} \quad \mathbf{P} = \frac{1-d}{N}(I - dM)^{-1}\mathbf{1}, \quad (8)$$

where I is the identity matrix of size $N \times N$. This algebraic method will be well-defined if $I - dM$ is invertible. Indeed, we have

Theorem 2.2 *Let I be the identity matrix and M be the adjacency matrix defined above. If $d \in (0, 1)$, then*

$$I - dM$$

is invertible.

Proof. We first recall a matrix $A = [a_{ij}]_{1 \leq i, j \leq N}$ is diagonally dominant if

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^N |a_{ij}|$$

for all $i = 1, \dots, N$. It is easy to see that $(I - dM)^\top$ is diagonally dominant matrix by using (6) with $d < 1$. It is known (cf. [27], p. 178) that every diagonally dominant matrix is nonsingular, and hence is invertible. So is $I - dM$. ■

Certainly, finding the inverse $(I - dM)^{-1}$ is expensive in computation. An easy approach is to use an iterative method, for example, power method. Starting with $\mathbf{P}_0 = \mathbf{1}/N$, we iteratively compute

$$\mathbf{P}_k = \mathcal{G}\mathbf{P}_{k-1} = \frac{1-d}{N}\mathbf{1} + dM\mathbf{P}_{k-1} \quad (9)$$

by using Theorem 2.1 for $k = 1, 2, \dots$. We now show that $\mathbf{P}_k, k \geq 1$ converge. Let us use the ℓ_1 norm for matrices to measure the errors $\mathbf{P}_k - \mathbf{P}$. Recall the ℓ_1 norm for matrix A with $A = [a_{ij}]_{1 \leq i, j \leq N}$ is

$$\|A\|_1 = \max_{j=1, \dots, N} \sum_{i=1}^N |a_{ij}|.$$

Then we have

Theorem 2.3 *Let \mathbf{P} be the PageRank vector defined by the algebraic method. For all $k \geq 1$,*

$$\|\mathbf{P}_k - \mathbf{P}\|_1 \leq Cd^k,$$

where C is a positive constant.

Proof. It is easy to see from (7) and the definition (9) of iterations that

$$\mathbf{P}_k - \mathbf{P} = dM(\mathbf{P}_{k-1} - \mathbf{P}) = \dots = (dM)^k(\mathbf{P}_0 - \mathbf{P}).$$

Thus, letting $C = \|\mathbf{P}_0 - \mathbf{P}\|_1$,

$$\|\mathbf{P}_k - \mathbf{P}\|_1 \leq d^k \|M^k\|_1 C \leq Cd^k \|M\|_1^k.$$

Finally we note that $\|M\|_1 = 1$ by using (6) to complete the proof. ■

We remark that the convergence of \mathbf{P}_k to \mathbf{P} was shown numerically in [37] for full size and a half size link databases. The rates of convergence are nearly the same for these two sizes. Also in [23], the contributor(s) of this wiki page uses the power method for the leading eigenvector to explain the convergence of the iterative method above without giving a convergence rate as it is based on the ratio of the second largest and the largest eigenvalues of the Google matrix. The second eigenvalue of the Google matrix was studied in [24] where the rate of convergence of PageRank was determined to be 0.85, the damping factor d . As mentioned in a previous section, the convergence rate has been determined by several methods already. Our result in Theorem 2.3 gave another proof. It is much simpler and easier fashion than the study in [24].

2.4 A New Local Update Scheme

We discuss how to do local updates. That is, in practice, webpages are updated constantly in the sense that the number $L(\mathbf{v}_i)$ of links from webpage \mathbf{v}_i changes very often, usually increases. A new adjacency matrix M is needed updated everyday. Instead of using an updated matrix M to compute a new PageRank vector \mathbf{P} for all components, we discuss an approach to update partial components of \mathbf{P} . We need the following well-known formula:

Lemma 2.1 (Sherman-Morrison's formula, 1949) *If A is invertible and \mathbf{x}, \mathbf{y} are two vectors such that $\mathbf{y}^\top A^{-1} \mathbf{x} \neq -1$, then $A + \mathbf{x} \mathbf{y}^\top$ is invertible and*

$$(A + \mathbf{x} \mathbf{y}^\top)^{-1} = A^{-1} - \frac{A^{-1} \mathbf{x} \mathbf{y}^\top A^{-1}}{1 + \mathbf{y}^\top A^{-1} \mathbf{x}}.$$

Assume that we have an updated $\ell_+(\mathbf{v}_i, \mathbf{v}_j)$ for a fixed integer j and M_+ be the updated adjacency matrix. Let $n = n(j)$ and $n_+ = n_+(j)$ be the previous and updated numbers $L(\mathbf{v}_j)$ of links from webpage \mathbf{v}_j . Let us focus the case that only more webpage links are added to the existing links from the webpage \mathbf{v}_j . In this case $n_+ > n$. Hence, we have

$$M_+ = M + \mathbf{x}_j \mathbf{v}_j^\top,$$

where $\mathbf{x}_j = [\ell_+(\mathbf{v}_i, \mathbf{v}_j) - \ell(\mathbf{v}_i, \mathbf{v}_j), i = 1, \dots, N]^\top$ with $\ell_+(\mathbf{v}_i, \mathbf{v}_j)$ is the updated version of $\ell(\mathbf{v}_i, \mathbf{v}_j)$ and

$$\|M - M_+\|_1 = 1 - n/n_+ + \frac{n_+ - n}{n_+} \leq 2 \frac{n_+ - n}{n_+}.$$

We shall use \mathbf{P}_+ to be the updated PageRank vector satisfying

$$\mathbf{P}_+ = \frac{1-d}{N} \mathbf{1} + d M_+ \mathbf{P}_+.$$

By Theorem 2.2, we know $I - d M_+$ is invertible. By the inverse formula in Lemma 2.1, we know $\mathbf{v}_j^\top (I - d M)^{-1} \mathbf{x}_j \neq -1$ and hence, letting $\alpha = 1 + \mathbf{v}_j^\top (I - d M)^{-1} \mathbf{x}_j \neq 0$,

$$(1 - d M_+)^{-1} = (1 - d M)^{-1} - (I - d M)^{-1} \mathbf{x}_j \mathbf{v}_j^\top (1 - d M)^{-1} / \alpha. \quad (10)$$

It follows

$$\begin{aligned}
\mathbf{P}_+ &= (1 - dM_+)^{-1} \frac{1-d}{N} \mathbf{1} \\
&= (1 - dM)^{-1} \frac{1-d}{N} \mathbf{1} - \frac{1}{\alpha} (I - dM)^{-1} \mathbf{x}_j \mathbf{v}_j^\top (1 - dM)^{-1} \frac{1-d}{N} \mathbf{1} \\
&= \mathbf{P} - \frac{1}{\alpha} (I - dM)^{-1} \mathbf{x}_j \mathbf{v}_j^\top \mathbf{P} \\
&= \mathbf{P} - \frac{P_j}{\alpha} (I - dM)^{-1} \mathbf{x}_j.
\end{aligned} \tag{11}$$

where P_j is the j^{th} component of vector \mathbf{P} . Let us write $\epsilon = \frac{P_j}{\alpha} (I - dM)^{-1} \mathbf{x}_j$ for the update for \mathbf{P} to obtain \mathbf{P}_+ and we next discuss how to find the update efficiently as solving $(I - dM)^{-1} \mathbf{x}_j$ is expensive in computation.

We need the following (cf. [27], p. 198)

Lemma 2.2 (Neumann Series) *If A is an $n \times n$ matrix and $\|A\|_1 < 1$, then $I - A$ is invertible and*

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k.$$

As shown before $\|M\|_1 = 1$ and $d \in (0, 1)$, letting $A = dM$ in Lemma 2.2 above, we have

$$(I - dM)^{-1} \mathbf{x}_j = \sum_{k=0}^{\infty} d^k M^k \mathbf{x}_j.$$

Algorithm 2.1 (A Single Update) *Choose an integer K large enough such that d^K is within a given tolerance. Start with $\mathbf{y} = \mathbf{x}_j$ and for $k = 1, \dots, K - 1$, we compute*

$$\mathbf{y} = dM\mathbf{y} + \mathbf{x}_j.$$

Then writing $\mathbf{y} = [y_1, \dots, y_N]^\top$, we let $\alpha = 1 + y_j$ and $\epsilon = (P_j/\alpha)\mathbf{y}$. Thus, by (11), $\mathbf{P}_+ \approx \mathbf{P} - \epsilon$ within the given tolerance.

The above algorithm leads to sequentially updating \mathbf{P} , i.e. update one webpage or one keyword or one document at a time. See Remark 5.1 for a parallel update scheme.

3 Sociological Analysis

Sociologists have studied the linkages, networks for a long time. Many concepts of centrality measures have been created. In [7], Bonscich proposed a measure of centrality $c(\alpha, \beta)$ with normalization factor α and parameter β which is a function of unit vectors of \mathbb{R}^N . For each unit vector, say, \mathbf{v}_i ,

$$c_i(\alpha, \beta) = \sum_j (\alpha + \beta c_j(\alpha, \beta)) R_{ij} \tag{12}$$

where R_{ij} is an entry of matrix relation \mathbf{R} which may not be symmetric and main diagonal elements are zero. In matrix notation

$$\mathbf{c}(\alpha, \beta) = \alpha(I - \beta\mathbf{R})^{-1}\mathbf{R}\mathbf{1}. \quad (13)$$

See page 1173 of [7]. If we use the adjacent matrix for \mathbf{R} and normalization factor $\alpha = (1 - d)/N$, we see that (12) is just (2) when $\beta = d$, the damping factor. If $\alpha = (1 - d)/N$ and $\mathbf{R}\mathbf{1} = \mathbf{1}$, we can see that (13) is the same as the equation on the right-hand side of (8). Using these two parameters, the PageRank is just $\mathbf{c}(\alpha, \beta)$. According to Bonacich, when \mathbf{R} is asymmetric, e.g., $\mathbf{R} = M$, $\mathbf{c}(\alpha, \beta)$ measures prestige if $\beta > 0$. This gives a sociological justification that the Google's PageRank does measure the important relevance scores for each webpage.

Note that when \mathbf{R} is symmetric, $\mathbf{c}(\alpha, \beta)$ measures centrality. Also, as in [7], β can be negative. For example, in a communication network,

As pointed out by Bonacich, his measure $\mathbf{c}(\alpha, \beta)$ seems hopelessly ambiguous.,

4 Improved Search Algorithms

We now explain the improvements discussed in the previous section in mathematical terminology. Recall all the notations and definitions in a previous section. We use page to denote a webpage, a keyword or phrase, a document, etc..

4.1 Convergence Analysis of In- and Out-Linkage Search Algorithm

For each page \mathbf{v} , $A_{\mathbf{v}}$ is the collection of all pages connected from \mathbf{v} . We now search all $\mathbf{u} \in A_{\mathbf{v}}$ to see if \mathbf{u} connects to \mathbf{v} also. We denote by $A_{\mathbf{v}}^+$ the collection of such page $\mathbf{u} \in A_{\mathbf{v}}$. Note that $A_{\mathbf{v}}^+$ could be an empty set if \mathbf{v} is an unknown page to the world. Anyway, $A_{\mathbf{v}}^+ \subset A_{\mathbf{v}}$ is a subset. Let

$$\tilde{L}(\mathbf{v}) = \#(A_{\mathbf{v}}) + \#(A_{\mathbf{v}}^+), \quad (14)$$

where $\#(A)$ stands for the number of entries in set A . It is called the cardinality of A . We define the new adjacency entry

$$\tilde{\ell}(\mathbf{v}, \mathbf{u}) = \begin{cases} 1/\tilde{L}(\mathbf{v}), & \text{if } \mathbf{u} \text{ is not connected to } \mathbf{v} \text{ from } \mathbf{u} \\ 2/\tilde{L}(\mathbf{v}), & \text{if } \mathbf{u} \text{ is connected to } \mathbf{v} \\ 0 & \text{otherwise.} \end{cases} \quad (15)$$

According to our Plan 1, the new PageRank function $\tilde{P}(\mathbf{v})$ is now defined by

$$\tilde{P}(\mathbf{v}) = d \sum_{\mathbf{u} \in A_{\mathbf{v}}} \tilde{P}(\mathbf{u})\tilde{\ell}(\mathbf{u}, \mathbf{v}) + \frac{1-d}{N}, \quad \forall \text{ any unit vector } \mathbf{v} \in \mathbb{R}^N, \quad (16)$$

where $d \in (0, 1)$ is a damping factor. To find the value of $\tilde{P}(\mathbf{v})$, we use similar iterative algorithm as in Section 2. Let

$$\tilde{M} = [\tilde{\ell}(\mathbf{v}_i, \mathbf{v}_j)]_{1 \leq i, j \leq N}, \quad (17)$$

We are now ready to define

Algorithm 4.1 (In- and Out-linkage Search Algorithm) Starting from $\tilde{\mathbf{P}}_0 = \mathbf{1}/N$, we iteratively compute

$$\tilde{P}_k = \frac{1-d}{N} \mathbf{1} + d\tilde{M}\tilde{P}_{k-1}, \quad \forall k = 1, 2, \dots \quad (18)$$

For convenience, we write $\tilde{\mathbf{P}}$ to the new PageRank vector of all pages $\mathbf{v}_i, i = 1, \dots, N$.

Based on the analysis in Section 2, we can conclude the following

Theorem 4.1 Let $\tilde{\mathbf{P}}$ be the PageRank vector defined above. For all $k \geq 1$,

$$\|\tilde{\mathbf{P}}_k - \tilde{\mathbf{P}}\|_1 \leq Cd^k, \quad \forall k \geq 1,$$

where C is a positive constant independent of k .

4.2 Convergence Analysis of Second Order linkage Search Algorithm

Next we study our improved search algorithm 2. In addition to use $A_{\mathbf{v}}^+$, we count the links from $\mathbf{u} \in A_{\mathbf{v}}$. That is, we look at $A_{\mathbf{u}}$. More number of links in $A_{\mathbf{u}}$, the less contribution to \mathbf{u} according to the sociological justification in the previous section. Thus, we define

$$n(\mathbf{v}) = \#(A_{\mathbf{v}}) + \sum_{\mathbf{u} \in A_{\mathbf{v}}} \frac{1}{\#(A_{\mathbf{u}})} \quad (19)$$

and

$$\hat{\ell}(\mathbf{u}, \mathbf{v}) = \begin{cases} \frac{1 + \frac{1}{\#(A_{\mathbf{u}})}}{n(\mathbf{v})}, & \text{if } \mathbf{u} \in A_{\mathbf{v}} \\ 0 & \text{otherwise.} \end{cases} \quad (20)$$

It is easy to see

Lemma 4.1 For any page \mathbf{v}

$$\sum_{j=1}^N \hat{\ell}(\mathbf{v}_j, \mathbf{v}) = 1. \quad (21)$$

Thus the new PageRank function $\hat{P}(\mathbf{v})$ is now defined by

$$\hat{P}(\mathbf{v}) = d \sum_{\mathbf{u} \in A_{\mathbf{v}}} \hat{P}(\mathbf{u}) \hat{\ell}(\mathbf{u}, \mathbf{v}) + \frac{1-d}{N}, \quad \forall \text{ any unit vector } \mathbf{v} \in \mathbb{R}^N, \quad (22)$$

where $d \in (0, 1)$ is a damping factor. To find the value of $\hat{P}(\mathbf{v})$, we do the iterations as before. Letting

$$\widehat{M} = \left[\hat{\ell}(\mathbf{v}_i, \mathbf{v}_j) \right]_{1 \leq i, j \leq N}, \quad (23)$$

Algorithm 4.2 (Second Order linkage Search Algorithm) Starting from $\hat{\mathbf{P}}_0 = \mathbf{1}/N$, we iteratively compute

$$\hat{P}_k = \frac{1-d}{N} \mathbf{1} + d\widehat{M}\hat{P}_{k-1}, \quad \forall k = 1, 2, \dots \quad (24)$$

Similarly, we write $\hat{\mathbf{P}}$ to the new PageRank vector of all pages $\mathbf{v}_i, i = 1, \dots, N$.

Based on the analysis in Section 2, we can conclude the following

Theorem 4.2 *Let $\hat{\mathbf{P}}$ be the PageRank vector defined above. For all $k \geq 1$,*

$$\|\hat{\mathbf{P}}_k - \hat{\mathbf{P}}\|_1 \leq Cd^k, \quad \forall k \geq 1,$$

where C is a positive constant independent of k .

5 Yahoo Search Engine

Yahoo search engine is quite different from the Google search engine conceptually and hence, computationally. The following description of the Yahoo search engine is based on [38].

Yahoo emphasizes on the similarity of any two webpages, two documents, or two web hubs. There are many sources of similarity information including link similarity, text similarity, multimedia component similarity, click through similarity, title similarity, URL similarity, cache log similarity, and etc.. For simplicity, let us consider the search engine for documents. Even for documents, there are many categories of similarity such as title similarity, author similarity, text similarity and etc.. For example, the text similarity is the dot product of two vectors which are the frequency usages of various words. That is, suppose that two documents use total M words. For one document, let \mathbf{x} be the vector of length M whose each entry is the frequency usage of a word in the document. Similar for vector \mathbf{y} for the other document. Then $\mathbf{x} \cdot \mathbf{y}$ denote the similarity between these two documents.

Suppose that the base set of all on-line documents contains N documents. The global similarity matrix $W = [w_{ij}]_{1 \leq i, j \leq N}$ with $w_{ij} \geq 0$ representing the strength of the similarity between document i and document j . Assume that $w_{ij} \leq M$ for all $i, j = 1, 2, \dots, N$. Suppose that there are K categories of various similarities. For each $k = 1, \dots, K$, let $\mathbf{x}(k) = [x_{1k}, x_{2k}, \dots, x_{Nk}] \in \mathbb{R}^N$ be a vector with nonnegative entries satisfying

$$\sum_{i=1}^N x_{ik} = 1.$$

It is a vector of scores or confidence values for all documents with respect to the k^{th} category of similarity. Note that the global similarity matrix combines all sources of similarity information together. Thus, $w_{ij} = f(w_{ij}(1), \dots, w_{ij}(K))$ with $w_{ij}(k), k = 1, \dots, K$ being the similarity between document i and document j with respect to the category k . Yahoo considers a global similarity objective function

$$P(\mathbf{x}) = \sum_{k=1}^K \mathbf{x}(k)^T W \mathbf{x}(k) = \sum_{k=1}^K \sum_{i,j=1}^N x_{ik} w_{ij} x_{jk}.$$

Yahoo engine will have to first do data preparation or preprocessing by generating the global similarity matrices W based on training and learning and then compute the matrix \mathbf{x} to maximize the objective function value by solving the following maximization problem:

$$\max_{\substack{\mathbf{x}(k) \in C \\ 1 \leq k \leq K}} P(\mathbf{x}), \quad (25)$$

where $\mathbf{x}(k) = [x_{1k}, x_{2k}, \dots, x_{Nk}]^\top$ and $C = \{(x_1, x_2, \dots, x_n), x_i \geq 0, i = 1, \dots, N, \sum_{i=1}^N x_i = 1\}$ is a convex set in \mathbb{R}^N . This solution matrix \mathbf{x} determines the search results according to the numerical values in the entries of \mathbf{x} . For document i , the document j is most similar to document i if the distance between vectors $x_{ik}, k = 1, \dots, K$ and $x_{jk}, k = 1, \dots, K$ is the smallest.

Since $P(\mathbf{x})$ is a quadratic function of \mathbf{x} and hence continuous, $P(\mathbf{x})$ achieves its maximum within the bounded set C . It is easy to see

Lemma 5.1 *Let $M = \max_{1 \leq i, j \leq n} w_{ij}$ be the maximum of the entries of the global similarity matrix. Then $P(\mathbf{x}) \leq MK$ for all $\mathbf{x}(k) \in C, k = 1, \dots, K$.*

Proof. Since $\sum_{i=1}^n x_i = 1$ and $x_i \geq 0$ for all $i = 1, \dots, N$, we have

$$P(\mathbf{x}) = \sum_{k=1}^K \sum_{ij} x_{ik} w_{ij}(k) x_{jk} \leq \sum_{k=1}^K \sum_{ij} M x_{ik} x_{jk} = \sum_{k=1}^K M \sum_{i=1}^n \sum_{j=1}^n x_{jk} x_{ik} = MK.$$

■

However, $P(\mathbf{x})$ may achieve its maximum at several vectors. For example, letting $K = 1, N = 3$ and W be the identity matrix of size 3×3 , we know the maximum is 1 when $\mathbf{x} = (1, 0, 0), (0, 1, 0), (0, 0, 1)$. Certainly, this is just a pathological example as no one uses an identity matrix for the global similarity matrix. The algorithm Yahoo uses produces a unique maximizer each time as it will be discussed below.

To solve the maximization problem, Yahoo uses the Growth Transformation algorithm—(GTA). The most important advantages of this algorithm are that it guarantees a monotonic increase sequence of the objective function values and the computation is stable as each step is a convex combination of the previous step. For simplicity, we assume that $K = 1$. In fact, Yahoo engine does a loop to go through each category using the following GTA.

Algorithm 5.1 (the Growth Transformation) *Starting with a trained vector $\mathbf{x}^{(1)}$ in C , for $k = 1, 2, \dots$, one computes*

$$\mathbf{x}_j^{(k+1)} = \frac{\mathbf{x}_j^{(k)} \frac{\partial}{\partial x_j} P(\mathbf{x}^{(k)})}{\sum_{i=1}^N \mathbf{x}_i^{(k)} \frac{\partial}{\partial x_i} P(\mathbf{x}^{(k)})} \quad (26)$$

for $j = 1, \dots, N$.

This algorithm is mainly based on an inequality proved in [5]. For convenience, we present a short version of the proof due to the degree $d = 2$.

Lemma 5.2 (Baum and Eagon, 1967) *Let $P(\mathbf{x})$ be a homogeneous polynomial of degree d with nonnegative coefficients. Let $\mathbf{x} \in C$ the convex defined above. For any $\mathbf{x} \in C$, let*

$$\mathbf{y} = (y_1, \dots, y_N) \text{ with } y_j = \frac{x_j \frac{\partial}{\partial x_j} P(\mathbf{x})}{\sum_{i=1}^N x_i \frac{\partial}{\partial x_i} P(\mathbf{x})}, j = 1, \dots, N.$$

Then $P(\mathbf{y}) > P(\mathbf{x})$ unless $\mathbf{y} = \mathbf{x}$.

Proof. For convenience, let us use $P(\mathbf{x}) = \mathbf{x}^\top W \mathbf{x}$ to prove this inequality, i.e., we assume that $d = 2$. Clearly, P is a homogeneous polynomial of degree $d = 2$. Since the entries of W are nonnegative, $P(\mathbf{x})$ satisfies the assumptions of this lemma. Note that

$$\frac{\partial}{\partial x_i} P(\mathbf{x}) = 2 \sum_{j=1}^n w_{ij} x_j$$

and

$$\sum_{i=1}^N x_i \frac{\partial}{\partial x_i} P(\mathbf{x}) = 2P(\mathbf{x}).$$

Thus, we first use Cauchy-Schwarz's inequality and then the geometric and arithmetic mean inequality to have

$$\begin{aligned} P(\mathbf{x}) &= \sum_{i,j=1,\dots,n} w_{ij} x_i x_j = \sum_{i,j=1}^n (w_{ij} y_i y_j)^{1/3} \left(w_{ij}^{2/3} x_i x_j \frac{1}{(y_i y_j)^{1/3}} \right) \\ &\leq \left(\sum_{i,j=1}^n w_{ij} y_i y_j \right)^{1/3} \left(\sum_{i,j=1}^n w_{ij} x_i x_j \left(\frac{x_i x_j}{y_i y_j} \right)^{1/2} \right)^{2/3} \\ &\leq \left(\sum_{i,j=1}^n w_{ij} y_i y_j \right)^{1/3} \left(\sum_{i,j=1}^n w_{ij} x_i x_j \frac{1}{2} \left(\frac{x_i}{y_i} + \frac{x_j}{y_j} \right) \right)^{2/3} \end{aligned}$$

Since $y_i = x_i \sum_{j=1}^n w_{ij} x_j / O(x)$, we have

$$\begin{aligned} \sum_{i,j=1}^n w_{ij} x_i x_j \frac{1}{2} \left(\frac{x_i}{y_i} + \frac{x_j}{y_j} \right) &= \frac{1}{2} \sum_{i,j=1}^n w_{ij} x_i x_j \left(\frac{P(\mathbf{x})}{\sum_{j=1}^n w_{ij} x_j} + \frac{P(\mathbf{x})}{\sum_{i=1}^n w_{ij} x_i} \right) \\ &= \frac{P(\mathbf{x})}{2} \left(\sum_{i=1}^n x_i \sum_{j=1}^n w_{ij} x_j / \left(\sum_{k=1}^n w_{ik} x_k \right) + \sum_{j=1}^n x_j \sum_{i=1}^n w_{ij} x_i / \left(\sum_{k=1}^n w_{kj} x_k \right) \right) \\ &= \frac{P(\mathbf{x})}{2} \left(\sum_{i=1}^n x_i + \sum_{j=1}^n x_j \right) = P(\mathbf{x}). \end{aligned}$$

since $\mathbf{x} \in C$. Combining the above discussion together, we have

$$P(\mathbf{x}) \leq \left(\sum_{i,j=1}^n w_{ij} y_i y_j \right)^{1/3} (P(\mathbf{x}))^{2/3}.$$

It thus follows that $P(\mathbf{x}) \leq \sum_{i,j=1}^n w_{ij} y_i y_j = P(\mathbf{y})$. Note that from the above discussion, the inequalities become equalities for Cauchy-Schwarz's inequality and geometric and arithmetic mean inequality when $\mathbf{y} = \mathbf{x}$. Therefore, we have a strictly inequality when $\mathbf{y} \neq \mathbf{x}$. We have thus completed the proof. ■

With this lemma, we are able to show the convergence of Algorithm 5.1. Due to the monotonically increasing property of $P(\mathbf{x}^{(k)})$ and the boundedness of $P(\mathbf{x})$ by Lemma 5.1, we see that $P(\mathbf{x}^{(k)})$ is convergent to $P(\mathbf{x}^*)$ for some point \mathbf{x}^* . In practice, the iterations may end in a finite steps. Note that for $\mathbf{x} \in C$, $\nabla P(\mathbf{x}) \neq 0$ and each component of $\nabla P(\mathbf{x})$ is strictly positive. Thus, objective function values are increasing until P achieves its maximum. In this case, by Lemma 5.2, we have $\mathbf{x}^{(n+1)} = \mathbf{x}^{(n)}$ for some integer n .

Also, $\mathbf{x}^{(k)}, k \geq 1$ are all in C and hence, are bounded. It follows that there exists a convergent subsequence. Due to the nonuniqueness of the maximizers, we can not conclude the convergence of the whole sequence.

Suppose that W is a positive definite, $P(\mathbf{x})$ is a convex function. when \mathbf{x} and \mathbf{y} defined in Lemma 5.2 are close enough, we have

$$\nabla P(\mathbf{x}) \cdot (\mathbf{y} - \mathbf{x}) \geq 0 \quad (27)$$

since $P(\mathbf{y}) > P(\mathbf{x})$. It follows that

$$P(\mathbf{y}) - P(\mathbf{x}) = P(\mathbf{y} - \mathbf{x}) + 2(\mathbf{y} - \mathbf{x})^\top W \mathbf{x} = P(\mathbf{y} - \mathbf{x}) + (\mathbf{y} - \mathbf{x})^\top \nabla P(\mathbf{x}) \geq \lambda_n \|\mathbf{y} - \mathbf{x}\|^2, \quad (28)$$

where λ_n denotes the smallest eigenvalue of W and $\|\mathbf{z}\|$ denotes the ℓ_2 norm of vector $\mathbf{z} \in \mathbb{R}^n$. Therefore we have

Theorem 5.1 *Suppose that W is symmetric and positive definite. Let M be the largest value of the entries of W . Let λ_n be the smallest eigenvalue of W . Let $\mathbf{x}^{(k)}, k \geq 1$ be the sequence from Algorithm 5.1. Then there exists a convergent subsequence, say $\mathbf{x}^{(m_k)}, k \geq 1$ which converge to a maximizer \mathbf{x}^* and satisfy*

$$\sum_{k=1}^{\infty} \|\mathbf{x}^{(m_{k+1})} - \mathbf{x}^{(m_k)}\|^2 \leq C < \infty,$$

where C is a positive constant dependent on λ_n and M .

Proof. We have discussed that the sequence $\mathbf{x}^{(k)}$ has a convergent subsequence. For convenience, let us say $\mathbf{x}^{(k)}$ converges to \mathbf{x}^* . Thus, $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$ will be very close when $k \geq K_0$ for a positive integer K_0 . Thus,

$$P(\mathbf{x}^{(k+1)}) - P(\mathbf{x}^{(k)}) \geq P(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \geq \lambda_n \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2$$

by using (28) and (27). Hence,

$$\lambda_n \sum_{k \geq K_0} \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|^2 \leq M - P(\mathbf{x}^{(K_0)}) \leq M < \infty.$$

This completes the proof. ■

Although Yahoo does not discuss much mathematics behind their generating the global similarity matrix W , one approach called *matrix completion* can be used. The matrix completion is a research topic in mathematics which is recently actively studied. It starts from the well-known Netflix problem. Indeed, Netflix (cf. [36]) made available publicly such a set of data with about 10^5 known entries of its movie rating matrix of size about 5×10^5 times 2×10^5 and challenged the research community to complete the movie recommending matrix with root mean square error less than 0.8563. This matrix completion problem has been studied actively ever since. We refer the following papers, [26], [41], and [29] for in-depth study and the references therein.

6 Remarks

We have the following remarks in order:

Remark 6.1 We will not discuss other link-based ranking systems for webpages such as the HITS algorithm invented by Jon Kleinberg (used by Ask.com), the IBM CLEVER project, IDD Information Services designed by Robin Li who found Baidu in China, the TrustRank algorithm and etc..

Remark 6.2 Next we discuss how to in parallel update the Ragerank simultaneously for multiple webpages and keywords and names of documents at a time. We need the following

Lemma 6.1 (Woodbury matrix identity) Suppose that A is invertible and C is also invertible. If $A + UCV$ is invertible, then

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

The above formula can be rewritten in the following form

Lemma 6.2 Suppose that U and V are unitary matrices and C is a diagonal matrix containing nonnegative values as entries. If $A + UCV$ is invertible, then

$$(A + UCV^\top)^{-1} = A^{-1} - (VC^\dagger U^\top A + I)^{-1}A^{-1}, \quad (29)$$

where C^\dagger is the pseudo-inverse of C .

Proof. First of all, we have

$$(A + UCV^\top)^{-1} = A^{-1} - A^{-1}U(C^\dagger + V^\top A^{-1}U)^{-1}V^\top A^{-1}$$

by using the same direct proof for Woodbury formula. Furthermore, we use the invertibility of U , V and A to rewrite the right-hand side of the above formula to get the equality in (29). ■

For the Google matrix $A = I - dM$ and $UCV^\top = -d \sum_{j=1}^k \mathbf{x}_j \mathbf{v}_j^\top$, we know most entries of matrix $\sum_{j=1}^k \mathbf{x}_j \mathbf{v}_j^\top$ of size $N \times N$ are zero. There is a nonzero block which is of size $n_k \times n_k$ with integer $n_k \ll N$ if k is not very big. For convenience, let us say the principal block of size $n_k \times n_k$ are nonzero. It is easy to find the singular value decomposition of this block, $-\sum_{j=1}^k \mathbf{x}_j \mathbf{v}_j^\top = U_1 C_1 V_1^\top$. Then let $U = \text{diag}(U_1, I_{N-n_k})$, $C = \text{diag}(C_1, \mathbf{0}_{N-n_k})$ and $V = \text{diag}(V_1, I_{N-n_k})$, where I_{N-n_k} is the identity matrix of size $N - n_k$ and $\mathbf{0}_{N-n_k}$ is zero block of size $N - n_k$. In general C_1 is not invertible. We can write it in the following form: if C_1 is of rank $r_k < n_k$,

$$C_1 = \begin{bmatrix} C_{11} & 0 \\ 0 & 0 \end{bmatrix},$$

where C_{11} is diagonal matrix containing all the nonzero singular values of $-\sum_{j=1}^k \mathbf{x}_j \mathbf{v}_j^\top$.

From (29), we need to find $(VC^\dagger U^\top A + I)^{-1}$. First, we have

$$VC^\dagger U^\top A = \begin{bmatrix} V_1 & 0 \\ 0 & I_{N-n_k} \end{bmatrix} \begin{bmatrix} \frac{1}{d} C_{11}^{-1} & 0 \\ 0 & \mathbf{0}_{N-n_k} \end{bmatrix} \begin{bmatrix} U_1^\top & 0 \\ 0 & I_{N-n_k} \end{bmatrix} \begin{bmatrix} I_{n_k} - dM_{11} & -dM_{12} \\ -dM_{21} & I_{N-n_k} - dM_{22} \end{bmatrix}.$$

for a matrix \tilde{C}_1 which is of size $n_k \times n_k$. Then we have

$$VC^\dagger U^\top A = \begin{bmatrix} \frac{1}{d} \tilde{C}_1 & 0 \\ 0 & \mathbf{0}_{N-n_k} \end{bmatrix} \begin{bmatrix} I_{n_k} - dM_{11} & -dM_{12} \\ -dM_{21} & I_{N-n_k} - dM_{22} \end{bmatrix},$$

where the Google matrix M is written in terms of the same division of blocks as $VC^\dagger U^\top$. Furthermore,

$$\begin{aligned} VC^\dagger U^\top A + I &= \begin{bmatrix} \frac{1}{d} \tilde{C}_1 (I_{n_k} - dM_{11}) + I_{n_k} & -\tilde{C}_1 M_{12} \\ 0 & I_{N-n_k} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{d} \tilde{C}_1 & 0 \\ 0 & I_{N-n_k} \end{bmatrix} \left(\begin{bmatrix} d\tilde{C}_1^{-1} - dM_{11} & -dM_{12} \\ 0 & \mathbf{0}_{N-n_k} \end{bmatrix} + I \right). \end{aligned}$$

Finally, we have

$$(VC^\dagger U^\top A + I)^{-1} = \left(I + d \begin{bmatrix} \tilde{C}_1^\dagger - M_{11} & -M_{12} \\ 0 & \mathbf{0}_{N-n_k} \end{bmatrix} \right)^{-1} \begin{bmatrix} d\tilde{C}_1^\dagger & 0 \\ 0 & I_{N-n_k} \end{bmatrix}.$$

It follows the updated PageRank vector is

$$\begin{aligned} \mathbf{P}_+ &= (I - dM - dUCV^\top)^{-1} \frac{1-d}{N} \mathbf{1} \\ &= (I - dM)^{-1} \frac{1-d}{N} \mathbf{1} - \left(I + d \begin{bmatrix} \tilde{C}_1^\dagger - M_{11} & -M_{12} \\ 0 & \mathbf{0}_{N-n_k} \end{bmatrix} \right)^{-1} \begin{bmatrix} d\tilde{C}_1^\dagger & 0 \\ 0 & I_{N-n_k} \end{bmatrix} (I - dM)^{-1} \frac{1-d}{N} \mathbf{1} \\ &= \mathbf{P} - \left(I + d \begin{bmatrix} \tilde{C}_1^\dagger - M_{11} & -M_{12} \\ 0 & \mathbf{0}_{N-n_k} \end{bmatrix} \right)^{-1} \begin{bmatrix} d\tilde{C}_1^\dagger & 0 \\ 0 & I_{N-n_k} \end{bmatrix} \mathbf{P}. \end{aligned}$$

We are now ready to present an algorithm for multiple updates.

Algorithm 6.1 (Multiple Updates) Choose an integer K large enough such that d^K is within a given tolerance. We first compute U_1, C_1, V_1 as explained above. Then compute \tilde{C}_1 . For the current PageRank vector \mathbf{P} , modify it to be

$$\tilde{\mathbf{P}} = \begin{bmatrix} d\tilde{C}_1^\dagger & 0 \\ 0 & I_{N-n_k} \end{bmatrix} \mathbf{P}$$

and then compute the following iterations

$$\mathbf{P}_+ \approx \mathbf{P} - \sum_{j=0}^K d^j \begin{bmatrix} \tilde{C}_1^\dagger - M_{11} & -M_{12} \\ 0 & \mathbf{0}_{N-n_k} \end{bmatrix}^j \tilde{\mathbf{P}}.$$

Similarly we can discuss how to compute M_+ efficiently when adding a new webpage/document/link to the database. We omit the detail here.

Remark 6.3 *The importance of webpages related to each keyword, document or webpage can also be calculated based on the number of hits from the world. However, this is very easily be scrolled up by some artificial hits.*

Acknowledgment: The authors would like to thank Dr. John Lavery for his encouragement and support which enable them to carry out the research in this paper.

References

- [1] T. Agryzkov, J. L. Oliver, L. Tortosa, J. F. Vicent, An algorithm for ranking the nodes of an urban network based on the concept of PageRank vector. *Appl. Math. Comput.* 219 (2012), no. 4, 2186–2193.
- [2] F. K. Andersson and S. D. Silvester, The Mathematics of Internet Search Engines, *Acta Appl. Math.* 104 (2008), 211–242.
- [3] K. Avrachenkov, N. Litvak, D. Nemirovsky, N. Osipova, Monte Carlo methods in PageRank computation: when one iteration is sufficient, *SIAM J. Numer. Anal.* 45 (2007) 890–904.
- [4] F. Backaker, The Google markov chain: convergence speed and eigenvalues, Master Thesis, Uppsala University, Sweden.
- [5] L. Baum and R. C. Eagon, An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology, *Bull. Amer. Math. Soc.*, 73(1967), pp. 360–363.
- [6] M. Benzi, E. Estrada, and C. Klymko, Ranking hubs and authorities using matrix functions, *Linear Algebra and its Applications*, 438(2013), 2447–2474.
- [7] P. Bonachich, Power and centrality: a family of measures, *Amer. J. Sociology*, 92(1987), 1170–1182.
- [8] C. Brezinski, M. Redivo-Zaglia, and S. Serra-Capizzano, Extrapolation methods for PageRank computations, *C. R. Acad. Sci. Paris, Ser I.* 340 (2005) 393–397.
- [9] C. Brezinski and M. Redivo-Zaglia, The PageRank vector: properties, computation, approximation and acceleration, *SIAM J. Matrix Anal. Appl.* 28(2006), 551–575.
- [10] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30: 107–117.
- [11] M. Brinkmeier, PageRank Revisited, *ACM Transactions on Internet Technology*, 6(2006), 282–301.
- [12] F. Chung and W. Zhao, A sharp PageRank algorithm with applications to edge ranking and graph sparsification. *Algorithms and models for the web graph*, 214, *Lecture Notes in Comput. Sci.*, 6516, Springer, Berlin, 2010.

- [13] F. Chung, A. Tsiatas, Finding and visualizing graph clusters using PageRank optimization, *Internet Math.* 8 (2012), no. 1-2, 46-72.
- [14] L. Eden, A Note on the Eigenvalues of the Google Matrix, *arXiv:math/0401177v1*, 2004.
- [15] D. Gleich, L. Zhukov and P. Berkhin, Fast Parallel PageRank: A Linear System Approach, Technical Report L-2004-038, Yahoo! Research Labs, 2004.
- [16] G. H. Golub and C. Greif, An Arnold type algorithm for computing PageRank, *BIT Numerical Mathematics*, 46(2006), 759–771.
- [17] G. H. Golub and C. D. Meyer, Using the QR factorization and group inverse to compute, differentiate, and estimate the sensitivity of stationary probabilities for Markov chains, *SIAM J. Algebra Discr.*, 17 (1986), pp. 273-281.
- [18] I. C. F. Ipsen, R. S. Wills, Mathematical properties and analysis of Google’s PageRank. *Bol. Soc. Esp. Mat. Apl. Se MA No. 34* (2006), 191-196.
- [19] [I.C.F. Ipsen, T.M. Selee, PageRank computation, with special attention to dangling nodes, *SIAM J. Matrix Anal. Appl.* 29 (2007) 1281–1296.
- [20] H. Ishii, R. Tempo, E.-W. Bai, Distributed randomized PageRank algorithms based on web aggregation over unreliable channels, in: *Proc. 49th IEEE Conf. on Decision and Control*, pp. 6602-6607, 2010.
- [21] H. Ishii, R. Tempo, E.-W. Bai, Distributed randomized PageRank algorithms over unreliable channels, in: L. Qiu, et al. (Eds.), *Developments in Control Theory towards Glocal Control*, Institution of Engineering and Technology, 2012, pp. 147-156.
- [22] H. Ishii, R. Tempo, and E. W. Bai, PageRank computation via a distributed randomized approach with lossy communication, *Systems & Control Letters* 61 (2012) 1221-1228.
- [23] <http://en.wikipedia.org/wiki/PageRank>
- [24] Haveliwala, T. and S. Kamvar. (2003). The Second Eigenvalue of the Google Matrix. Stanford University Technical Report: 7056. *arXiv:math/0307056*.
- [25] S. Kamvar, T. Haveliwala, G. Golub, Adaptive methods for the computation of PageRank, *Linear Algebra Appl.* 386 (2004) 51-65.
- [26] R.H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.
- [27] Kincaid, D. and W. Cheney. (2002). *Numerical Analysis: Mathematics of Scientific Computing*, Brooks/Cole.
- [28] J. Kleinberg, Authoritative sources in a hyperlinked environment, *J. ACM* 46 (1999) 604-632.
- [29] M. J. Lai, Yin, W. T. and Xu, Y. Y., On unconstrained nonconvex minimizations for sparse vector and low-rank matrix recovery, accepted for publication in *SIAM J. Numer. Analysis*, (2013).

- [30] A. N. Langville and C. D. Meyer, Deeper inside PageRank, *Internet Math.*, 1 (2005), pp. 335-380.
- [31] A. N. Langville and C. D. Meyer, A survey of eigenvector methods for Web information retrieval, *SIAM Rev.*, 47 (2005), pp. 135-161.
- [32] A. N. Langville and C. D. Meyer, Updating Markov chains with an eye on Google's PageRank, *SIAM J. Matrix Anal. Appl.* 27(2006), 968–987.
- [33] A. Longville and C. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton Univ. Press, 2006.
- [34] Y. Lin, X. Shi, Y. Wei, On computing PageRank via lumping the Google matrix, *Journal of Computational and Applied Mathematics* 224 (2009) 702–708.
- [35] R. Nath and N. Kumar, To Overcome HITS Rank Similarity Conflicion of Web Pages using Weight Calculation and Rank Improvement, *Amer. Inst. Physics Conf. Proc.* 1414 (2011), 77 doi: 10.1063/1.3669935.
- [36] Netflix Prize at <http://www.netflixprize.com/>.
- [37] Page, L., (1997). "PageRank: Bringing Order to the Web", Stanford Digital Library Project, talk. August 18, 1997 (archived 2002).
- [38] M. Palmer, G. Sun, and H. Zha, United States Patent No. 6,990,628 B1, Filed on June 14, 1999 and Date of Patent Jan. 24, 2006.
- [39] S. Serra-Capizzano, Jordan canonical form of the Google matrix: a potential contribution to the PageRank computation, *SIAM J. Matrix Anal. Appl.*, 27(2005), 305–312.
- [40] J. Wu and Y. Tan, PageRank algorithm optimization and improvement, *Computer Engineering and Application*, 45(2009), 56–59.
- [41] Z. Wen, W. Yin, and Y. Zhang. Solving a low-rank factorization model for matrix completion by a nonlinear successive over-relaxation algorithm. Rice University CAAM Report TR10-07, 2010.

A Summary of Sociological Concepts Related to Social Network and Its Techniques for Quantifying Social Cohesion, Social Position, Social Distance

Dawn Robinson, Dept. of Sociology, University of Georgia. Athens, GA 30602

January 16, 2013

1 Background

We are moving rapidly from a society built around relationships in homes, neighborhoods, workplaces, places of worship, and voluntary organizations, to a globally connected society with interactions that span large spatial and social distances. Sociological understanding of this transformation has yet to be achieved. This transition can be thought of as a transition from strong to weak ties. Humans form direct communities online, in the context of social networking sites, email communications, and virtual worlds, etc. They also, however, form indirect communities via the development of shared tastes, consumption patterns, media access, etc. Understanding the character and consequences of these direct and indirect relationships has been a key focus of business such as Google, Amazon, Pandora, and Yahoo, but has been relatively under-examined by contemporary computational social science.

In the spring of 2011, access to social networking sites was largely hailed as a prime facilitator of individuals across North Africa and the Middle East as they organized expressions of social unrest and political discontent on a massive scale and then communicated the results of those organized expressions with lightening speed. On the civilian side, the marketing community is keenly interested in exploiting sociological links, especially the weak sociological links that have recently become emplaced in and accessible through the Internet. Worldwide, internet usage is increasing at an astounding rate, particularly the use of social networking sites. The number of adult internet users in the United States doubled between 2008 and 2010 (Hampton, Goulet, Rainie Purcell 2011[19]). A recent Pew Research Center survey (reported that in the United States, social networking site use has risen from 26% of all adults in 2008 to 47% in 2010 (Rainie, Purcell, Smith 2011[31]). Most of these users (92%) are on Facebook; 13% use Twitter (Rainie, Purcell, Smith 2011). Evidence from this report suggests that the movement from organizationally and geographically organized communities to online communities has augmented rather than supplanted other types of sociality. Internet users are even more likely than the average American to belong to voluntary groups or organizations (80% versus 56%). Social networking site users are even more likely than other internet users to belong to organized groups, with Twitter users being the most likely to belong to voluntary groups or organizations. Facebook users are also more politically engaged than other U.S. adults (Hampton, Goulet, Rainie Purcell 2011). So, while mediated interactions are taking place at unprecedented rates, they do not seem to be supplanting other, more conventional forms of social organization. Rather, it is likely that these types of relations interact with one another in ways that are not yet adequately understood.

2 Useful Sociological Concepts

2.1 Tie Strength and System Size

Social network techniques for analyzing structure and position within social systems largely developed to understand strong sociological links (families, hierarchical/command organizations, communities with specific structure, nation states, etc.). These techniques were later applied to the study of weak social ties (acquaintances, occasional encounters, etc.), but there has been relatively little comparison of the differences between the two types of social systems. Strong sociological links are responsible for important aspects of human society, including but not limited to the ability to build entities that allow for coordinated action of large numbers of people. Strong sociological links involve loyalty and giving/accepting orders due to monetary and religious relationships. Weak sociological links, on the other hand, are the most common sources of new information which are now increasingly available to collect via internet. Both strong and weak links play positive roles in human society and in human progress. However, with the Internet and many other networking capabilities, the balance between strong and weak links is changing. As the accessibility of all links increases, the relative importance of the weak links is increasing.

Tie strength is a property of relationships, and generally refers to the intensity of commitment, constraint, or emotion attached to a particular link. Another way to characterize the difference between conventional social environments and those in the world of Web 2.0 is in terms of system size and complexity. While not equivalent, these are somewhat conflated in the scholarly literature. So, when we talk about research on weak versus strong ties, we primarily are distinguishing between analysis of the relational structure of small, often bounded groups, versus analysis of large, complex social systems.

An advantage of studying small, bounded groups is the ability to work with whole networks. Many of the approaches to looking at social structure and social position in the social network literature rely on graph theory and utilize the entire matrix of relations in their computation (Carrington, Scott, and Wasserman 2005). Much of the focus in the contemporary social network environment is on very large scale, sparse, and complex social systems. It remains an open question whether social network measurement approaches developed to understand small whole networks are optimal for understanding larger, sparser, more complex communication networks or knowledge networks based on weak links. In fact, some recent research suggests that other dynamics whose properties we thought we understood in simpler networks may operate differently in more complex systems. In light of this, we will survey the social network literature regarding the measurement and quantification of various indicators of social capital including social cohesion, social position, and social distance and consider to their potential applicability to large complex systems of weak relations.

2.2 Social Cohesion

Social cohesion has long been a subject of investigation in the study of groups (Albert 1953, Cartwright 1968, Lott 1961, Van Bergen and Koekebakker 1959). Social cohesion refers to the degree of solidarity within a group or social system and usually is defined as the degree of attraction and/or commitment toward the group/system held by individual members of the group/system. This premise of an individual-to-group relation has been the subject of some debate (can individuals actually relate to groups or do they only relate to other individuals?). The literature largely supports the idea that individuals can, indeed, have relationships with abstract groups and that these relationships can precede and supercede relations between individuals within those groups (for a review, see Friedkin 2004[?]).

Social network researchers also have used a variety of structural properties of the group to characterize its cohesion, rather than relying on individual reports of attraction or commitment to the group. These include the extent of positive ties within the group (Cartwright 1968[?]), the degree of symmetry among positive ties within the group (Moreno and Jennings 1937[?]), and the density of interpersonal relations (Festinger et al. 1950[?]). There are also numerous ways of identifying cohesive subgroups within larger systems (Wasserman and Faust 1994). These approaches, however, fail to capture the individual-to-group aspect of social cohesion, and run the risk of creating a tautology if we want to use structural features to predict social cohesion. Research using more classic measures of social cohesion reflecting individual commitment to the group finds that even groups that are large, complexly differentiated, sparse, and composed of weak ties can be highly cohesive when they have conducive structures (Doreian and Fararo 1998). These include reachability (Markovsky 1998[24]) and a low density of negative or punishing ties (Friedkin 2003[?]).

Additional research suggests that it rather the repeated activation of positive ties that produces group cohesion (Lawler, Thye and Yoon 2000; McPherson and Smith-Lovin 2002), rather than simply the quantity of them. This finding that ongoing nature of social relations differentiates them from other kinds of ties considered in isolation is related to the embeddedness approach in economic sociology (e.g., Granovetter 1985; Uzzi 1999). Embeddedness takes into account the degree to which mutual ties, reachability, and other opportunities for feedback loops in a network system create additional pressures toward trust and cohesion. When a relationship is embedded within a larger system of relationships, there is both a larger shadow of the past and a larger shadow of the future. This has the effect of creating more enduring relationships and a greater commitment to the groups or systems in which these relationships are embedded.

In summary, social cohesion is understood as a way of characterizing the stability and intensity of the relationship between individual members of a group or system and the group itself. Structural position (reviewed further below) and embeddedness predict the individual-to-group relationships from which group level cohesion derives. Structural conditions like density, reachability, reciprocity, and repetition of tie activation predict cohesion at the group level. While nodal reach and system reachability are not easily calculated in large or incomplete systems, reciprocity and repetition of tie activation are network features that can be accessed with egocentric data from sampled nodes, and so are features that might easily be used in understanding the dynamics of massive, complex systems.

2.3 Various Measures of Centrality

The mostly widely investigated social network measures are those characterizing social position sometimes called social prominence or network centrality. In the social network literature, there are four primary means of characterizing the structural position of a particular node degree, betweenness, closeness, and Bonacich power. These are methods of determining the centrality of a vertex within a graph. In the context of social networks, they are typically used to determine the importance of a particular person within the group or system. The usefulness of each of these methods of characterizing structural position depends on features of the system, the social context, and the nature of the resource flowing through the graph or network. We will describe these further below.

Degree. The simplest and among the most frequently used measure of structural position is degree based centrality. In symmetric networks, this is simply the number of lines or edges connecting a particular node or vertex to other nodes (Freeman 1979[16]). In simple affiliation systems, this is considered to be a basic characterization of popularity. A count of ones Facebook friends is a fairly ubiquitous contemporary

measure of degree centrality among contemporary college students. While Facebook friendships occur in a symmetric social network, many naturally occurring networks are fundamentally asymmetric in nature. Liking, respect, information seeking, and assistance are routinely exchanged asymmetrically. In these cases, it helps to distinguish between in-degree (number of ties received) versus out-degree (number of ties sent) to determine social prominence in a network. In-degree is a more precise measure of prominence when the resource flow is positive and or deferent (e.g., respect, advice-seeking). Out-degree may be a more precise measure of social prominence when the resource flow is the diffusion of information. Nodes with high out-degree can serve as gatekeepers in a social system. Asymmetry between out-degree and in-degree can also serve as a measure of node prominence. When group size is known, actor degree can be standardized by the group size in order to compare prominence of actors across groups (Wasserman and Faust 1994:179).

An elaboration of this approach developed by Bonacich (1987) uses iterative simultaneous equations to converge on an estimate of power that combines degree of actor with information about the actors relational neighborhood. This method recognizes that being connected to others with many connections can increase an actors importance in a positively connected (contagious) network and simultaneously decrease ones power in a negatively connected (competitive) network. Imagine a network in which Sally and Bob each have five friends. Sallys friends each have a high number of other friends; Bobs friends are isolates and are not friends with many others. If the social process of interest is contagious, like the diffusion of information, the transfer of disease, or even diffusion of positive regard, then Sally would have more influence than Bob. She gets her power by being connected to other highly connected others. In contrast, if the resource flowing across the network is competitive, then Bob may gain more power by being connected to others who are more dependent upon him for the competitive resource. Bonacichs algorithm accordingly allows for specifying the level and direction of attenuation in the network.

Closeness. A different approach to quantifying an actor's power in a social network is closeness-based centrality (Freeman 1979[16]). This approach presumes that a node's power is a function of it geodesics. The simplest version is simply an inverse of the sum of all distances from an actor to all other actors in a network. This approach to power is especially useful in positively connected (contagious) networks across which resources diffuse with some moderate rate of decay. An elaboration of the closeness centrality approach, called reach centrality, considers what portion of the network an actor can reach with each additional number of steps (Hanneman and Riddle 2005[20]).

Betweenness. A third general approach to quantifying an actor's power in a social network relies on betweenness-based centrality (Freeman 1977[15]). This approach recognizes that interactions between unconnected members of a network often critically depend on other actors in the system - especially those who lie on the paths between the two. The simplest measure of betweenness centrality simply counts all of the geodesics between all pairs of actors in a system which contain a particular actor. An elaboration of this idea, called information centrality, generalizes this to include all paths between all actors, weighted by the inverse of their lengths when calculating centrality (Stephenson and Zelen 1989[37]). This takes into account the idea that information does not always flow along the shortest path, and that actor can gain importance by controlling the flow across many paths, as well as by controlling only a few short paths.

Eigenvector Centrality. A fourth approach to quantifying structural position uses a factor analytic procedure to discount closeness to small local subnetworks (Bonacich 1972[2]). This approach, called eigenvector centrality, allows researchers to differentiate between proximity in the global structure and proximity in more local substructures, by computing principal components of the actor distance measures and generating an eigenvalue for each actor on each structural dimension (Freeman 1979). This

approach generates estimates of structural position that are very close to degree based centrality when there is a fairly flat distribution of degree, or in core-periphery structures, where high degree nodes are connected primarily to other high degree nodes (Bonacich 2007[?]). But, in structures with many connections between low degree and high degree nodes, the kinds of hierarchical clustering that characterizes most naturally occurring systems (Barabasi et al. 1999; Watts et al. 2002; Watts 2004 [39]) this method produces distinctively different and much more useful predictions.

Dependence-based power. Markovsky and colleagues (Markovsky et al. 1994; 1998 [23] and [24]) have developed another set of measures specifically designed to measure potential for structural influence in negatively connected, or competitive networks. The simplest form of such competition is when there is no resource flow (perfect decay) and nodes are limited to single exchange partners. In such networks, the structure of relationships in one part of system can constrain power relations within dyads far away in the system in systematic ways. The simplest of these graph-theoretic power indices (called GPI) subtracts the number of disadvantageous paths (the number of unique even length paths between a node and all other nodes in the system) from the advantageous paths (the number of unique odd length paths between a node and all other nodes) to generate a measure of dependence-based power. Individuals with many ties to individuals with no other ties (paths of length 1) are rewarded for their ties to dependent others. These measures have been used to successfully predict power use and exchange outcomes in a number of experimental studies of human interaction.

These measures of structural position have varying utility by context. In whole networks we can easily see how the decay rate of a resource flow determines which measure of structural position is most useful. When a resource flow is very quickly consumed, then degree based centrality may be the most useful. We can think of many social behaviors and affects that are like this. If Sally smiles at Bob, her smile is consumed. Bob can smile at another co-worker but it won't be Sally's smile. In this case, we want to understand patterns of single-transfer social behaviors and affects, degree based centrality may be sufficient. When a resource flow can travel through one or more nodes, but with some decay factor, closeness-based centrality becomes more important. Information flows across a network, but tends to lose veracity and sometimes change (increase or decrease) intensity with each transfer. Consequently, being closer to the source provides one with more accurate information. Being connected to many well-connected others may facilitate more influence than simply having the most friends. When individuals serve as liaisons across various densely connected regions in a system, they accrue betweenness-based power, allowing them to serve as gatekeepers for resources that have slower decay functions.

Closeness-based centrality, betweenness-based centrality, and dependence-based require whole networks for computation. This limits their utility to contexts where full information is available and raises the question their computational efficiency when applied to very large, sparse networks. Degree based centrality has the benefit of being easy to access and not requiring whole network information and so can easily be used in the context of egocentric data or in large, complex, and sparse networks. Some of the variants of degree based centrality, including eigenvector centrality and the variant, PageRank, used by Google's search engine, can be calculated with only the use of first-order ties, making them computationally much more efficient in large, sparse, and incomplete networks. With the consideration one additional order of relationships (out to 2nd order ties), the formula could be substantially improved to take into account the kind of local-distal effects better captured by whole-network approaches such as closeness, betweenness, and especially dependence-based power.

2.4 Social Distance

Social distance measures have been used in sociology for nearly a century. Many of these measures, like the classic Bogardus social distance scale (1925), are self-report attitudinal measures. There also is a long history, however, of quantifying social distance and social positions using social network techniques. Most of these techniques for identifying social distance at the dyadic level are the same as those used to identify social cohesion at the group or system level. The simplest of these techniques require that the geodesic distances between the nodes in a network or subnetwork be small. Others compare within-group ties to out-group ties. Other approaches make use of clustering or multi-dimensional scaling techniques to represent social distance along a small number of dimensions for visual interpretation.

Some of these techniques focus on identifying sets of structurally equivalent actors. One such approach uses Pearsons correlation as a measure of structural equivalence and uses the convergence of iterated correlations between relations as a means of partitioning into subsets. Using connection through music tastes as an example let us describe this classic technique (Breiger et al. 1975, White et al. 1976). Take an adjacency matrix A of actors and music purchases. Multiply the matrix A by its transpose A^* to get an actor X actor matrix of people connected through their shared music preferences. Correlate the rows and columns of this new matrix. Replace the values in the matrix with the results. Repeat until the cells are filled with 1s and 0s. Separate the 1's and 0's into separate matrices. Replace with values from original Actor by Actor matrix and start again. Each successive split will group actors into subgroups with more similarly shared patterns of relations to others (through shared music preferences). This creates a binary tree of partitions among actors, with all actors being partitioned into exhaustive and mutually exclusive subsets. Finally, partition the original actors by the resulting positions and permute the matrix to reveal the relationships between the structurally equivalent blocks.

A variant on this technique relies instead on the Euclidean distances between the ties to and from two actors, instead of using Pearsons correlations to capture degree of similarity (Burt 1976). For each pair of actors i and j take the Euclidean distance between rows i and j and columns i and j . When two actors are structurally equivalent (connected to the same other actors), they distance between them will be 0. Once a matrix of equivalence relations has been computed, actors can be partitioned into cohesive subgroups through the use of hierarchical clustering.

The correlation based measure of structural equivalence has the advantage of capturing the equivalence of actors who have similar kinds of relations with similar kinds of others, while the Euclidean distance measures

For example "long ties" (those connecting actors who are socially distant along other social dimensions) increase diffusion rates in some classic studies, like Granovetter's (1974) study of job seekers in which individuals were more likely to find jobs via weak (and long) ties than via strong ties. In other words, long ties speed simple contagions. Long ties, however, actually slow diffusion of information when adoption requires multiple affirmations. In these systems, long ties slow complex contagion processes (Centola and Macy 2007; Centola, Eguilez and Macy, 2007).

3 Conclusion

The PI and co-PI have immersed in the mathematical and sociological literatures on social networks and made some initial connections between them. Above, we have briefly summarized the We have summarized the sociology of social structure, position and influence in strong and weak networks. In the

appended document, we summarize the mathematical study about the Google search algorithm along with some suggested improvements based on and this sociological literature. To further explore and develop these connections and their implications will require a greater time investment than afforded by this seed project.

With more time, this research team could more deeply digest the existing information, and propose some new algorithms, making use of sociological insights to improve mathematically the characterization of relations in large, complex systems. The PI and co-PI are more than willing to continue our joint work toward a better understanding of the searching algorithms and discovering how to better situate them in the larger contexts of existing mathematical and sociological knowledge.

References

- [1] Bainbridge, W.S., Brent, E.E., Carley, K.M., Heise, D.R., Macy, M.W. Markovsky, B. and Skvoretz, J. 1994. Artificial social intelligence. *Annual Review of Sociology* 20:407–436.
- [2] Bonacich, Phillip. 1972. "Factoring and weighting approaches to status scores and clique identification." *Journal of Mathematical Sociology* 2(1): 113–120.
- [3] Bonacich, Phillip. 1987. "Power and centrality: A family of measures." *American journal of sociology* : 1170–1182.
- [4] Bonacich, Phillip, and Paulette Lloyd. 2001. "Eigenvector-like measures of centrality for asymmetric relations." *Social Networks* 23(3): 191–201.
- [5] Breiger, R.I., S. Boorman, and P. Arabie. 1975. An algorithm for clustering relational data with applications to social network analysis. *Journal of Mathematical Psychology* 12:329–383.
- [6] Burt, R.S. 1976. Positions in networks. *Social Forces* 55: 93–122.
- [7] Carley, K.. 1991. A Theory of Group Stability. *American Sociological Review*. 56:331-354. Centola, D. and Macy, M. 2007. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology* 113:702–734.
- [8] Carrington, Peter J., John Scott, and Stanley Wasserman. 2005. *Models and Methods in Social Network Analysis*. Cambridge University Press.
- [9] Centola, Damon and Michael Macy. 2007. Complex Contagions and the Weakness of Long Ties. *American Journal of Sociology* 113:702-734.
- [10] Centola, D. Eguiluz, V.M. and Macy, M. 2007. Cascade Dynamics of Complex Propagation *Physica A* 374: 449–456
- [11] Coleman, J. S, 1964. *An Introduction to Mathematical Sociology*. Free Press.
- [12] Coleman, J. S, 1990. *Foundations of Social Theory*. Harvard University Press.
- [13] Edling, C. R. 2002. Mathematics in Sociology, *Annual Review of Sociology*.

- [14] Feld, S.L. 1997. Mathematics in Thinking about Sociology. *Sociological Forum*. 12:3-9.
- [15] Freeman, L.C. 1977. A Set of Measures of Centrality based on Betweenness. *Sociometry*. 40: 35– 41
- [16] Freeman, L.C. 1979. Centrality in social networks: Conceptual clarification, *Social Networks*, 1: 215–39.
- [17] Gilbert, Nigel; Troitzsch, K. (2005). *Simulation and social science*. Simulation for Social Scientists (2 ed.). Open University Press.
- [18] Granovetter, M.S. (1973), The strength of weak ties, *Amer. J. Sociology* 78, 1360-1380.
- [19] Hampton, K.H., Goulet, L.S., Rainie, L., and Purcell, K.. 2011. *Social Networking Sites and Our Lives*. Pew Internet & American Life Project, June 16, 2011. <http://pewinternet.org/Reports/2011/Technology-and-social-networks.aspx>. Accessed June 18, 2011.
- [20] Hanneman, R. and Riddle, M. 2005. *Introduction to social network methods*. Riverside, CA: University of California, Riverside (published in digital form at <Http://faculty.ucr.edu/~hanneman/>)
- [21] Heckathorn, D. D. 1990. Collective sanctions and compliance norms: a formal theory of group-mediated social control. *Am. Sociol. Rev.* 55:366–84.
- [22] Kitts, J. A. 2006. *Simulation Modelling Practice and Theory*. 14:407–422.
- [23] B. Markovsky and E. J. Lawler. 1994. A New Theory of Group Solidarity. *Advances in Group Processes*. 11:113–37.
- [24] B. Markovsky, 1998. *Social Conceptions of Solidarity*. In *The Problems of Solidarity*. Gordon and Breach Publishers.
- [25] McPherson, J.M. and Ranger-Moore, J.R. 1991. Evolution on a Dancing Landscape: Organizations and Networks in Dynamic Blau Space. *Social Forces* 70:19-42
- [26] Mitchell, T.M. 2009. Mining Our Reality. *Science*. 326:1644-1645
- [27] J. Moody and D. R. White. 2004. Structural Cohesion and Embeddedness: A Hierarchical Concept of Social Groups. *American Sociological Review* 68(1):103-127.
- [28] Moreno, J. L. 1951. *Sociometry, Experimental Method and the Science of Society. An Approach to a New Political Orientation*. Beacon House, Beacon, New York.
- [29] Opsahl, T.; Agneessens, F.; Skvoretz, J. (2010). Node centrality in weighted networks: Generalizing degree and shortest paths. *Social Networks* 32: 245.
- [30] Orbach, T.L., House, J.S., Mero, R.P. and Webster, P.S. 1996. Marital Quality Over the Life Course. *Social Psychology Quarterly*. 59:162–171,
- [31] Rainie, L., Purcell, K., and Smith, A. 2011. *The Social Side of the Internet*. Pew Internet & American Life Project, June 16, 2011. <http://pewinternet.org/Reports/2011/The-Social-Side-ofthe-Internet.aspx>. Accessed June 18, 2011

- [32] Rashevsky, N.: 1947/1949 (2nd ed.). *Mathematical Theory of Human Relations: An Approach to Mathematical Biology of Social Phenomena*. Bloomington, ID: Principia Press.
- [33] Robinson, D.T. 1996. Identity and Friendship: Affective Dynamics and Network Formation. *Advances in Group Processes* 13:93–113
- [34] Robinson, D.T. 2007. Control Theories in Sociology. *Annual Review of Sociology*. 33:157–174.
- [35] Robinson, Dawn T., and James W. Balkwell.1995. "Density, transitivity, and diffuse status in task-oriented groups." *Social Psychology Quarterly* 58: 241–254.
- [36] Schneider, A. and Heise, D.R.. Simulating symbolic interaction. *Journal of Mathematical Sociology*, 20 1995: 271–287.
- [37] Stephenson, K. A. and Zelen, M., 1989. Rethinking centrality: Methods and examples. *Social Networks* 11, 1–37.
- [38] B. Uzzi, 1996. "The Sources and Consequences of Embeddedness for the Economic Performance of Organizations: The Network Effect." *American Sociological Review* 61: 674-98.
- [39] D. J. Watts, 2004. *Small Worlds*. Princeton, NJ: Princeton University Press.
- [40] D. J. Watts, P. S. Dodds, and M. E. J. Newman. 2002. Identity and Search in Social Networks. *Science* 296 (5571):1302–1305.
- [41] White, H. C., Boorman, S. A., and Breiger, R.L. 1976 Social structure from multiple networks: I. Blockmodels of roles and positions. *American Journal of Sociology* 81:730–780.